On Corruption-Robustness in *Performative* Reinforcement Learning

Vasilis Pollatos[1*], **Debmalya Mandal**[2†], **Goran Radanovic**[3]

[1] Archimedes/Athena RC, Greece
[2]University of Warwick
[3]Max Planck Institute for Software Systems
v.pollatos@athenarc.gr, debmalya.mandal@warwick.ac.uk, gradanovic@mpi-sws.org

Abstract

In *performative* Reinforcement Learning (RL), an agent faces a policy-dependent environment: the reward and transition functions depend on the agent's policy. Prior work on performative RL has studied the convergence of repeated retraining approaches to a *performatively* stable policy. In the finite sample regime, these approaches repeatedly solve for a saddle point of a convex-concave objective, which estimates the Lagrangian of a regularized version of the reinforcement learning problem. In this paper, we aim to extend such repeated retraining approaches, enabling them to operate under corrupted data. More specifically, we consider Huber's ϵ-contamination model, where an ϵ fraction of data points is corrupted by arbitrary adversarial noise. We propose a repeated retraining approach based on convex-concave optimization under corrupted gradients and a novel problem-specific robust mean estimator for the gradients. We prove that our approach exhibits last-iterate convergence to an approximately stable policy, with the approximation error linear in $\sqrt{\epsilon}$. We experimentally demonstrate the importance of accounting for corruption in performative RL.

Introduction

In *performative reinforcement learning* (Mandal, Triantafyllou, and Radanovic 2023; Rank et al. 2024), the learner operates in a policy-dependent environment, where the reward and transition functions are influenced by the learner's policy. A canonical example of such a setting is an RL system involving human users, such as RL-based recommender systems or chatbots: the RL policy affects user preferences, which in turn alters user engagement and behavior, thereby modifying the RL environment.

For an offline learner, *performativity* represents a specific type of distribution shift. Similar to standard reinforcement learning (RL), where the learner's policy influences the data it encounters due to the sequential decision-making process, performativity introduces an additional layer of complexity. The learner not only affects the observed data but also alters the underlying data generation process through its policy choices. This dual influence complicates the learning process, impacting the learner's ability to generalize effectively.

To alleviate this challenge prior work places sensitivity assumptions, formalized as Lipschitz conditions, which limit the extent the learner's policy can influence the reward and transition functions of the underlying environment. These conditions enable convergence guarantees to a performatively stable policy during repeated retraining. This process involves optimizing a regularized RL objective after each deployment round to update the current policy. In the finite data regime, the retraining step solves for a saddle point of a convex-concave objective, which approximates the Lagrangian of the regularized reinforcement learning problem.

In this paper, we additionally consider a third type of distribution shift, specifically induced by an adversary capable of corrupting training data. This scenario is particularly relevant in practical settings where performativity plays a role. For example, recommendation systems are vulnerable to Sybil attacks, while chatbots can be susceptible to poisoning poisoning attacks (e.g., see the Tay.ai incident (Neff 2016)). Therefore, it is crucial to explore strategies for achieving corruption-robustness in performative environments.

Our goal is to extend algorithmic and convergence results from prior work on perfomrative RL by allowing an ϵ fraction of data points in training data to be corrupted by adversarial noise—commonly referred to as the *Huber ϵ-contamination model*. We recognize two main challenges:

- First, this involves solving a convex-concave optimization problem where access to a given objective function is corrupted. For instance, gradient-based methods, which are commonly used for convex-concave optimization, rely on first-order gradient oracles to access the objective. However, existing guarantees for these methods typically assume that the outputs of gradient oracles are clean and not adversarially corrupted.
- Second, generic robust estimators, such as those used for estimating gradients, do not account for the problem structure, often making them impractical in challenging high-dimensional settings.

Contributions: Our work aims to resolve these challenges within the performative RL framework. Our contributions are as follow:

- We propose a robust version of Optimistic Optimistic

[*]This work was done as part of a research immersion lab at the Max Planck Institute for Software Systems.

[†]The work was done while at the Max Planck Institute for Software Systems.

Follow the Regularized Leader (OFTRL) for optimizing convex-concave objectives under corruption and provide theoretical analysis of thereof. This analysis shows that the robust OFTRL achieves optimal convergence rates and information-theoretically optimal terminal error. These results are of an independent interest and complement the concurrent work on gradient-based algorithms for convex-concave optimization under corrupted gradients. See the related work section for more details.

- We extend the existing algorithmic approach to performative reinforcement learning, making it robust against corruption. More specifically, we propose a novel repeated retraining approach based on robust OFTRL and a novel coordinate-wise robust mean estimator for estimating the gradients. The robust mean estimator is particularly suited for performative RL. We theoretically analyze our repeated retraining approach, showing that it exhibits last-iterate convergence to an approximately stable policy, with the approximation error that scales linearly with the square root of the corruption level ϵ.
- Using a simulation-based experimental testbed, we showcase the importance of accounting for corrupted gradients, and the efficacy of our approach.

The extended version of the paper (Pollatos, Mandal, and Radanovic 2024) provides additional information, including the proofs of our formal results and implementation details for the experimental analysis.

Related Work

We recognize several lines of works related to this paper: *performative prediction and RL*, *corruption-robust offline RL*, and *convex-concave optimization*, and *convex optimization under corrupted gradients*.

Performative Prediction and RL: Performative prediction models data distribution shifts due to the model deployment (Perdomo et al. 2020). Much prior work has studied convergence properties of algorithms to different solution concepts, including performative stability (Perdomo et al. 2020; Mendler-Dünner et al. 2020) and performative optimality (Izzo, Ying, and Zou 2021; Miller, Perdomo, and Zrnic 2021). In recent years, other variants of the canonical setting have been proposed, including multi-player variants (Narang et al. 2023; Piliouras and Yu 2023), variants that consider more nuanced state-dependent distribution shifts (Brown, Hod, and Kalemaj 2022; Li and Wai 2022), or variants that introduce constraints (Yan and Cao 2024) or bilevel optimization (Lu 2023). The closest to our setting is the work of (Mandal, Triantafyllou, and Radanovic 2023), who introduced the concept of performativity in reinforcement learning. The performative RL framework relates to Stackelberg stochastic games (Letchford et al. 2012; Zhong et al. 2021), in which a principal agent commits to a policy to which follower agents respond—making the principal's effective environment performative. However, the performative RL framework abstracts away from game-theoretic considerations, as it does not model performativity through game-theoretic agents. (Rank et al. 2024) extends the performative RL framework by considering gradual environment shifts, akin to those considered in the performative precision settings of (Brown, Hod, and Kalemaj 2022; Li and Wai 2022). Our paper contributes to the extensive literature on performative prediction (Hardt and Mendler-Dünner 2023) by introducing corruption-robustness to performative RL.

Corruption-Robust Offline RL: From a technical perspective, our results build on the analysis (Mandal, Triantafyllou, and Radanovic 2023), who adapted the minimax optimization problem (Zhan et al. 2022) for the offline setting of performative RL. Hence, our work relates to the vast literature on corruption robustness in offline RL (Zhang et al. 2022; Wu et al. 2022; Ye et al. 2024; Nika et al. 2024; Mandal et al. 2024). However, these works do not utilize corruption-robust minimax optimization in their frameworks, nor do their underlying framework model performativity effects. An additional discussion of how the bounds in some of these works compares to ours is provided in the convergence analysis of our approach.

Convex-Concave Optimization: Our approach relies on convex-concave optimization. There has been a vast literature on this topic, from the *fictitious play* algorithm (Robinson 1951) and the extra-gradient method (Korpelevich 1976; Tseng 1995) for solving bilinear optimization problems, to Gradient Ascent Descent-based algorithms for the general convex-concave optimization problem (Nemirovski 2004; Nesterov 2007; Tseng 2008; Nedić and Ozdaglar 2009; Mokhtari, Ozdaglar, and Pattathil 2019, 2020). Most relevant to our setting are works on convex-concave optimization that assume inexact oracles. (Juditsky, Nemirovski, and Tauvel 2011; Huang and Zhang 2022) consider biased stochastic gradient oracles. In (Beznosikov, Sadiev, and Gasnikov 2020), a zeroth-order biased oracle with stochastic and bounded deterministic noise is assumed and (Dvinskikh et al. 2022) consider a zeroth-order oracle corrupted by adversarial noise. Our work differs from these results as our setting has a bounded gradient corruption (bias) and we provide guarantees for the actual error of our algorithm instead of the expected. Arguably, the closest related work on convex-concave optimization to ours is the concurrent work of (Zhang et al. 2024). They provide convergence guarantees under adversarial noise for smooth convex-concave functions, but achieve this through a different algorithm. Hence, our results on convex-concave optimization complement theirs. Our information-theoretic lower bound on the duality gap has a different flavor than the lower bounds in (Zhang et al. 2024); the latter focuses on notions related to algorithmic *reproducibility*. Hence, our lower bound and the analysis we used for deriving it is novel.

Convex Optimization under Corrupted Gradients: In convex optimization gradient corruption has been more extensively studied (Polyak 1987; d'Aspremont 2008; Devolder 2013; Devolder, Glineur, and Nesterov 2014). From this line of research most relevant are the works of (Prasad et al. 2020) and (Wang, Mianjy, and Arora 2021), who study corruption in gradient descent due to poisoning attacks with applications to machine learning, as well as works of (Ahn et al. 2022), that study reproducibility in optimization in the face of inexact gradients.

Preliminaries

In this section, we provide the necessary background. Our main approach builds on convex-concave optimization under corrupted gradients. More specifically, we consider smooth convex-concave objectives. We propose a robust version of Optimistic Follow the Regularized Leader (OFTRL), and provide a convergence guarantee for it. In particular, we show that it exhibits $O(1/T)$ convergence rate for the duality gap. This result and the analysis are of independent interest, complementing prior work that studied convex-concave analysis under gradient corruption.

Notation. In the following sections, we use $\|\cdot\|$ to denote the L_2 norm $\|\cdot\|_2$, $[N]$ to denote the set $\{1, 2, ...N\}$ and $\Pi_{\mathcal{X}}(x)$ to denote the projection of a vector x on a set $\mathcal{X}$.

Convex-concave optimization

We consider the following constrained minimax problem

$$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} f(x, y), \tag{1}$$

where f is a convex-concave function, and $\mathcal{X}$ and $\mathcal{Y}$ are convex bounded domains. We denote the upper bounds to the radius of domains $\mathcal{X}$ and $\mathcal{Y}$ by D_X and D_Y respectively, i.e., $\max_{x\in\mathcal{X}} \|x\| \leq D_X$ and $\max_{y\in\mathcal{Y}} \|y\| \leq D_Y$. We assume that gradients of this function are bounded, i.e., $\max_{x\in\mathcal{X},y\in\mathcal{Y}} \|\nabla_x f(x, y)\| \leq G_X$ and $\max_{x\in\mathcal{X},y\in\mathcal{Y}} \|\nabla_y f(x, y)\| \leq G_Y$, for some constants G_X, G_Y. These assumptions are satisfied in the objective function that we will consider in the sections on perfomrative RL.

Contamination Model

We are interested in designing a robust optimization method for finding a saddle point of f which only has access to f through noisy (first order) gradient oracles with bounded noise norm. We assume we are given a sampling procedure that generates unbiased gradient samples an $1-\epsilon$ fraction of the time and adversarial samples (potentially with unbounded corruption) otherwise. This is a strong contamination model known as Huber contamination. We aim to design an optimization method that is robust in the sense that: a) its convergence guarantees have an information theoretically optimal dependence on the noise norm of the inexact gradient oracle, matching our lower bound in 2 and b) it deploys robust gradient estimators that trim the unbounded corruption down to a bounded error.

Smooth Convex-Concave Objectives

The convex-concave objective that we will study in the sections on performative RL satisfies smoothness. Hence, in this section, we consider a class of convex-concave functions f that are smooth. In particular, we assume that for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ functions $\nabla_x f(\cdot, y), \nabla_x f(x, \cdot), \nabla_y f(\cdot, y)$ and $\nabla_y f(x, \cdot)$ have Lipschitz constants L_{XX}, L_{XY}, L_{XY} and L_{YY} respectively, w.r.t. ℓ_2 norm. Similarly to the setting of exact gradient oracles, these smoothness conditions enable us to achieve a better convergence rate.

Robust OFTRL. To find a saddle point of f, we propose a robust version of gradient-based Optimistic Follow the Regularized Leader, as defined in Algorithm 1. The algorithm follows standard OFTRL steps (e.g., see (Orabona 2019)). It alternates between updating x_t and y_t, in optimizing for each a regularized objective. Regularizer ψ_X (resp. ψ_Y) is λ_X (resp. λ_Y) strongly convex and bounded over $\mathcal{X}$ (resp. $\mathcal{Y}$). For instance $\psi_X(x)$ could be equal to $\frac{\lambda_X}{2}\|x\|^2$ and $\psi_Y(y)$ could be equal to $\frac{\lambda_Y}{2}\|y\|^2$. Importantly, the algorithm utilizes robust gradient estimates in line 7. While adversarial samples can potentially have unbounded corruption, prior work has shown that there exist robust estimators for the Huber ϵ-contamination model (e.g. (Diakonikolas, Kane, and Pensia 2020)) that can filter the samples and guarantee that the gradient estimation has a bounded error w.r.t. the true gradient and this error scales with ϵ. For the analysis in the next subsection, we will assume black box access to a (robust) gradient estimation oracle upon query on some point (x_t, y_t). For the results on performative RL, we provide a problem-specific robust estimator.

Algorithm 1: Robust OFTRL

1: Initialize $\alpha, b, c > 0$
2: $(\lambda_X, \lambda_Y) \leftarrow 3(L_{XX} + L_{XY}\alpha + b,\ L_{YY} + L_{XY}/\alpha + c)$
3: $(g_{X,0},\ g_{Y,0}) \leftarrow (\mathbf{0}, \mathbf{0})$
4: **for** $t = 1, ...T$ **do**
5: $\quad x_t \leftarrow \arg\min_{x\in\mathcal{X}}\ \psi_X(x) + \langle g_{X,t-1}, x\rangle + \sum_{i=1}^{t-1}\langle g_{X,i}, x\rangle$
6: $\quad y_t \leftarrow \arg\min_{y\in\mathcal{Y}}\ \psi_Y(y) + \langle g_{Y,t-1}, y\rangle + \sum_{i=1}^{t-1}\langle g_{Y,i}, y\rangle$
7: $\quad$ Calculate robust estimations $g_{X,t}$ and $g_{Y,t}$ of $\nabla_x f(x_t, y_t)$ and $-\nabla_y f(x_t, y_t)$ respectively
8: **end for**
9: $(\bar{x},\ \bar{y}) \leftarrow (\frac{1}{T}\sum_{t=1}^{T} x_t,\ \ \frac{1}{T}\sum_{t=1}^{T} y_t)$
10: **Return** $\bar{x}, \bar{y}$

Analysis of Robust OFTRL

Next, we analyze the convergence guarantees of Robust OFTRL. Denoting the errors of robust gradient estimation as $\zeta_t^X = g_{X,t} - \nabla_x f(x_t, y_t)$ and $\zeta_t^Y = g_{Y,t} - \nabla_y f(x_t, y_t)$, we obtain the following result.

Theorem 1. *The output $(\bar{x}, \bar{y})$ of Algorithm 1 satisfies for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$:*

$$f(\bar{x}, y) - f(x, \bar{y}) \leq \frac{\psi_X(x) + \psi_Y(y)}{T} + \frac{\|\nabla_x f(x_1, y_1)\|^2}{\lambda_X \cdot T}$$
$$+ \frac{\|\nabla_y f(x_1, y_1)\|^2}{\lambda_Y \cdot T} + \frac{6D_X}{T}\sum_{t=1}^{T}\|\zeta_t^X\| + \frac{6D_Y}{T}\sum_{t=1}^{T}\|\zeta_t^Y\|.$$

The proof of this theorem is based on the results from (Orabona 2019), which consider the exact gradients. We see that the algorithm converges to an approximate saddle point with the $\frac{1}{T}$ convergence rate and the approximation error, i.e., asymptotic duality gap, dependent on the errors of robust gradient estimation. The $\frac{1}{T}$ convergence rate is optimal for smooth convex-concave problems, as shown in prior

work (Ouyang and Xu 2021). Next, we show that the dependence of the asymptotic duality gap on the gradient noise and domain radius is information theoretically optimal.

Let $\mathcal{A}$ be some deterministic algorithm that estimates saddle points. $\mathcal{A}$ has only access to a noisy gradient oracle of f that can be called T times. At timestep t the algorithm chooses a point (x_t, y_t) and the oracle returns $g_x(x_t, y_t) = \nabla_x f(x_t, y_t) + \zeta_t^X$ and $g_y(x_t, y_t) = \nabla_y f(x_t, y_t) + \zeta_t^Y$. The noise is bounded as follows: $\|\zeta_t^X\| \leq Z_X$ and $\|\zeta_t^Y\| \leq Z_Y \; \forall t \in [T]$. The algorithm has knowledge of the constants Z_X and Z_Y but does not know the exact noise values. In this setting we can derive the following lower bound:

Theorem 2. *Consider a deterministic algorithm $\mathcal{A}$ that estimates saddle points of convex concave functions $f(x, y)$ over the domain $\|x\| \leq D_X$, $\|y\| \leq D_Y$, where x and y are d-dimensional vectors, using T adaptive queries on noisy gradient oracles with $\|\zeta_t^X\| \leq Z_X$ and $\|\zeta_t^Y\| \leq Z_Y$ for all $t \in [T]$ and $Z_Y \leq D/2$, $Z_X \leq D/2$, where $D = min\{D_X, D_Y\}$. For any such $\mathcal{A}$:*

There exists a convex concave (bilinear) function $f(x, y)$ and a noise sequence realisation, such that $\mathcal{A}$ returns a point (x_0, y_0) that has distance at least $\frac{Z_X + Z_Y}{\sqrt{2}}$ from any saddle point of f and duality gap $f(x_0, y) - f(x, y_0) \geq \frac{1}{4} Z_Y D_Y + \frac{1}{4} Z_X D_X$ for some pair (x, y) inside the domain $\|x\| \leq D_X$, $\|y\| \leq D_Y$.

Theorem 1 and Theorem 2 provide a rather complete characterization of the convergence properties of robust OFTRL under corrupted gradients. To the best of our knowledge, these characterization results are novel (see the related work section for comparison to prior work). Moreover, we can prove similar results for Optimistic Mirror Descent Ascent, which is known to have an exponential last iterate convergence rate for objectives that satisfy the *Metric Subregularity* (MS) condition (Wei et al. 2021). We provide this analysis in the appendix of the extended version of the paper. The latter results are novel and of independent interest for minimax optimization and they could be useful for performative RL if we prove that the Lagrangian (3) satisfies the MS condition or if we simply make it satisfy MS by adding regularization on both variables.

Formal Setting

The focus of this work is on the performative reinforcement learning framework, introduced by (Mandal, Triantafyllou, and Radanovic 2023).

Policy-dependent Markov Decision Process

The performative reinforcement learning framework considers a policy-dependent Markov Decision Process (MDP) defined as tuple $M(\pi) = (\mathcal{S}, \mathcal{A}, P^\pi, r^\pi, \rho, \gamma)$, where: $\mathcal{S}$ is a finite state space, $\mathcal{A}$ is a finite action space, $\pi : \mathcal{S} \to \mathcal{P}(A)$ is a stochastic policy, $P^\pi : \mathcal{S} \times \mathcal{A} \to \mathcal{P}(\mathcal{S})$ is a transition function, with $P^\pi(s, a, s')$ denoting the probability of transition to state s' when action a is taken in state s, $r^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function, $\rho \in \mathcal{P}(\mathcal{S})$ is the initial state distribution, and $\gamma \in [0, 1)$ is the discount factor. We denote $S = |\mathcal{S}|$ and $A = |\mathcal{A}|$, and we assume that rewards are bounded, i.e., $r^\pi(s, a) \leq R$ for some (unknown) constant R.

Performatively Stable Policy

To define a solution concept in this framework, we define the value of policy π in $M(\pi')$ given initial state distribution ρ as $V_{\pi'}^\pi(\rho) = \mathbb{E}_\tau \left[\sum_{t=0}^\infty \gamma^t \cdot r^{\pi'}(s_t, a_t) | \rho\right]$, where $\tau = (s_0, a_0, s_1, a_1, ...)$ is a trajectory obtained by executing policy π in MDP $M(\pi')$. The solution concept of interest is a *performatively stable* policy π_S, which satisfies: $\pi_S \in \arg\max_\pi V_{\pi_S}^\pi(\rho)$.

Occupancy Measures

We denote by $d^\pi(s, a)$ the occupancy measure of policy π in MDP $M(\pi)$, i.e., $d^\pi = \mathbb{E}_\tau \left[\sum_{t=0}^\infty \gamma^t \cdot \mathbb{1}\left[s_t = s, a_t = a\right] | \rho\right]$. Occupancy measure d^π satisfies the Bellman flow constraint

$$\forall s : \rho(s) + \gamma \sum_{s', a} d^\pi(s', a) P^\pi(s', a, s) = \sum_a d^\pi(s, a).$$

For a generic $d \geq 0$, we define policy $\pi^{\downarrow d}$ as

$$\pi^{\downarrow d}(s, a) = \begin{cases} \frac{d(s,a)}{\sum_{a'} d(s,a')} & \text{if } \sum_{a'} d(s, a') > 0 \\ \frac{1}{A} & \text{othw.} \end{cases}. \quad (2)$$

If d is a valid occupancy measure in $M(\pi^{\downarrow d})$ (i.e., if it satisfies the Bellman flow constraints), the occupancy measure of $\pi^{\downarrow d}$, i.e., $d^{\pi^{\downarrow d}}$, is equal to d. In general, $d^{\pi^{\downarrow d}}$ and d may differ. The occupancy measure of a performatively stable policy is denoted by d_S.

Data Generation Process

We are interested in a finite sample, offline RL regime. The data generation process is assumed to be i.i.d.: (s_i, a_i) is sampled from normalized d_n, i.e., $(s_i, a_i) \sim (1 - \gamma) \cdot d_n$, $r_i = r_n(s_i, a_i)$ and s_{i+1} is sampled from P_n transition kernel, i.e., $s_{i+1} \sim P_n(s_i, a_i, \cdot)$. We also make a coverage assumption that $d_n(s, a)$ is positive for all $s \in \mathcal{S}, a \in \mathcal{A}$. This assumption can be satisfied if the dynamics make all states reachable and we mix some exploratory random policy with π_n. In particular, we assume that if $d(s, a) \geq c$ then $\forall (s, a) \in \mathcal{S} \times \mathcal{A}$, $d^{\pi^{\downarrow d}}(s, a) \geq B(c) > 0$. For the rest of this work, all deployed policies will satisfy this condition and thus the coverage assumption will hold with constant $B(c)$.

Contamination Model

We consider the Huber ϵ-contamination model, where an ϵ fraction of data points can be corrupted. In this case, both transition and reward can be corrupted. In particular, when the adversary corrupts a sample (s_i, a_i, s_i', r_i), next state s_i' can be replaced by any $s_c \in \mathcal{S}$ and reward r_i can be placed by any $r_c \in \mathbb{R}$. We assume that s_i and a_i are not corrupted. The latter assumption is made for technical simplicity. In the appendix of the extended version of the paper, we show how the assumption could be avoided and still derive similar results with a more complicated analysis.

Corruption-Robust Performative RL

We follow prior work on performative RL, and study repeated retraining and its convergce to an approximate performatively stable point. In repeated retraining, the policy is

retrained after each deployment round. In a canonical setting, we have access to the MDP model $M(\pi_n)$, where π_n is the policy deployed in round n. We will denote this MDP model by $M_n = M(\pi_n)$, and its reward and transition function by r_n and P_n, respectively. Assuming access to M_n, repeated retraining optimizes the following regularized RL objective after each round n:

$$d^*_{n+1} \in \arg\max_{d \geq \mathcal{C}(M_n)} \sum_{s,a} d(s,a) \cdot r_n - \frac{\lambda}{2} \cdot \|d\|_2^2 ,$$

where $\mathcal{C}(M_n)$ is the space of occupancy measures compatible with the MDP M_n, satisfying:

$$\forall s : \rho(s) + \gamma \sum_{s',a} d(s',a) P_n(s',a,s) = \sum_a d(s,a).$$

Note that π_n is obtained from d^*_n, and is defined as $\pi_n := \pi^{\downarrow d^*_n}$. To directly optimize from data, Mandal, Triantafyllou, and Radanovic (2023) consider the following minimax optimization problem:

$$d^*_{n+1}, h^*_{n+1} \in \arg\max_{d \geq 0} \arg\min_h \mathcal{L}(d, h, M_n), \qquad (3)$$

where objective $\mathcal{L}$ is the Lagrangian of the regularized RL objective, defined as:

$$\mathcal{L}(d,h,M_n) = -\frac{\lambda}{2}\|d\|_2^2 + \sum_s h(s)\rho(s) + \sum_{s,a} d(s,a) \times \left[r_n(s,a) - h(s) + \gamma \sum_{s'} P_n(s,a,s') h(s') \right].$$

Given dataset D_n containing m samples (s_i, a_i, s'_i, r_i), we can replace $\mathcal{L}$ with its empirical version $\hat{\mathcal{L}}$:

$$\hat{\mathcal{L}}(d,h,M_n) = -\frac{\lambda}{2}\|d\|_2^2 + \sum_s h(s)\rho(s) + \sum_{(s,a,r,s') \in D_n} \frac{d(s,a)}{d_n(s,a)} \cdot \frac{r - h(s) + \gamma \cdot h(s')}{m \cdot (1-\gamma)},$$

where m is the size of D_n and $d_n(s,a)$ is the occupancy measure of policy π_n in M_n. This allow us to directly optimize from data and establish last-iterate convergence guarantees from finite samples.

Remark 1. *In the above framework we assumed for simplicity knowledge of the occupancy measure d_n that generates the samples. In practice, we can estimate it up to arbitrary accuracy from the samples using Monte-Carlo.*

Robust Repeated Retraining

We build on the repeated retraining approach described above, but consider the case where dataset D_n is corrupted, according to the corruption model specified in the formal setting. Our approach is depicted in Algorithm 2.

In each round n, the algorithm first collects a contaminated data D_n by deploying policy π_n in $M(\pi_n)$. The next step is to approximately solve problem (3) – given that D_n is corrupted directly utilizing $\hat{\mathcal{L}}$ instead of $\mathcal{L}$ may not yield any guarantees. Hence, we apply robust OFTR, described in the preliminaries, with $f = \mathcal{L}$ and the robust gradient estimators from the next subsection. Finally, the algorithm calculates new policy π_{n+1} by mixing the occupancy measure obtained via robust OFTRL with an exploratory random policy and applying Eq. (2). We do this by adding some positive constant c to $\bar{d}$. The mixing step ensures the coverage property for the next iteration will be satisfied.

In the next subsection, we first propose a novel robust gradient estimator of $\mathcal{L}$, which can be combined with robust OFTRL to approximately solve (3). We provide guarantees on the estimation error for this estimator that scales with the corruption level ϵ. We then focus on the convegence analysis of Algorithm 2, and show that it exhibits last-iterate convergence to an approximately stable policy, with the approximation error proportional to $\sqrt{\epsilon}$.

Algorithm 2: Robust Repeated Retraining

```
π_0 = 0
for n = 1, ..., N do
    D_n ← Sample d^{π_n} + Huber ε-contamination
    d̄_{n+1} ← Apply Robust OFTRL with f = L and gradient estimators from Section Robust Gradient Estimation on D_n
    d̃_{n+1} ← d̄_{n+1} + c, where c > 0
    π_{n+1} ← π^{↓ d̃_{n+1}}, where π^{↓d} is defined in Eq. (2).
end for
Return d̃_N
```

Robust Gradient Estimation

We now focus on robust estimation of gradients $g_d := \nabla_d \mathcal{L}(d,h,M_n)$ and $g_h := \nabla_h \mathcal{L}(d,h,M_n)$. If some of the samples are corrupted, naive averaging may not suffice. Therefore, we explore robust alternatives. We propose the following steps:

- For the gradient w.r.t. d, given a subset $\{(s_i, s'_i, a_i, r_i) \mid i \in [\tilde{m}]\}$ of D_n, with corruption level ϵ, we apply a robust mean estimator to the dataset $D_d := \{\hat{g}^i_d \mid i \in [\tilde{m}]\}$, where each sample $\hat{g}^i_d$ is a single-entry $|\mathcal{S}| \cdot |\mathcal{A}|$-dimensional vector constructed by sample (s_i, s'_i, a_i, r_i) according to the formula $\hat{g}^i_d(s,a) = \mathbb{1}[(s_i,a_i) = (s,a)] \cdot \frac{\gamma h(s'_i) - h(s_i) + r_i}{(1-\gamma)\cdot d_n(s_i,a_i)}$. D_d contains both corrupted and clean samples. Each clean sample $\hat{g}^i_d$ is an unbiased estimator of $g_d + \lambda d$. Finally, we add $-\lambda d(s,a)$ to the robust mean of D_d.
- For the gradient w.r.t. h, given a subset $\{(s_i, s'_i, a_i, r_i) \mid i \in [\tilde{m}]\}$ of D_n (disjoint with that used for g_d), with corruption level ϵ, we apply a robust mean estimator to the dataset $D_h := \{\hat{g}^i_h \mid i \in [\tilde{m}]\}$, where each sample $\hat{g}^i_h$ is a $|\mathcal{S}|$-dimensional vector constructed by sample (s_i, s'_i, a_i, r_i) according to the formula $\hat{g}^i_h(s') = d(s_i,a_i)\frac{\gamma \mathbb{1}[s'=s'_i] - \mathbb{1}[s'=s_i]}{(1-\gamma)\cdot d_n(s_i,a_i)}$. Each clean sample $\hat{g}^i_h$ is an unbiased estimator of $g_h - \rho$. Finally, we add ρ to the robust mean of D_h.

Algorithm 3: Robust coordinate-wise mean

Input: $\{\hat{g}_d^i \mid i \in [\tilde{m}]\}$, ϵ
for $k = 1, ..., S \cdot A$ **do**
  $data \leftarrow \{\hat{g}_d^1[k], ..., \hat{g}_d^{\tilde{m}}[k]\}$
  $Med_k \leftarrow median(data)$
  $clean \leftarrow (1-\epsilon) \cdot \tilde{m}$ closest $data$ entries to Med_k
  $\hat{g_d}[k] \leftarrow mean(clean)$
end for
Return $\hat{g_d}$

Robust Mean Estimators. As a robust mean estimator we consider any estimator whose error has one (statistical) term, vanishing with the number $\tilde{m}$ of samples and one bias term polynomial to the frequency ϵ of corrupted samples. The error of a robust estimator should not scale with the magnitude of corruption, especially when corruption can be unbounded, as it can happen in the reward samples in our setting. For the estimation of g_d, we use Algorithm 3. For the estimation of g_h it suffices to apply naive averaging to achieve the kind of result that we wish. The errors of gradient estimators are analysed in the following theorem.

Theorem 3. *Let us use $\hat{g}_h = \frac{1}{\tilde{m}} \sum_{i=1}^{\tilde{m}} \hat{g}_h^i$ to estimate g_h and Algorithm 3 to estimate g_d, and assume that the corruption level in the respective datasets is bounded by $\epsilon < 0.5$. Then with probability at least $1-\delta$ the estimation errors satisfy the following guarantees:*

$$\|\hat{g}_h - g_h\|_1 \leq \underbrace{\frac{4}{(1-\gamma)^2 B(c)} \left(\frac{\sqrt{S \, \log(4S/\delta)}}{\sqrt{\tilde{m}}} + \epsilon \right)}_{E_1(\tilde{m}, \epsilon, \delta)},$$

$$\|\hat{g}_d - g_d\|_2 \leq \underbrace{6\sqrt{SA} \frac{2h_{max} + R}{(1-\gamma)B(c)} \left(\frac{\sqrt{2 \log\left(\frac{4SA}{\delta}\right)}}{\sqrt{\tilde{m}}} + 2\epsilon \right)}_{E_2(\tilde{m}, \epsilon, \delta)}.$$

Convergence Analysis

Our goal is to show that the repeated optimization approach as specified in Algorithm 2, outputs a solution that is approximately stable. We restrict the domain of variables d and h in Algorithm 2 as follows: $\mathcal{D} = \{d : 0 \leq d(s,a) \leq \frac{1}{1-\gamma}\}$ and $\mathcal{H} = \{h : -h_{max} \leq h(s,a) \leq h_{max}\}$, where $h_{max} > 0$. Furthermore, we will assume that D_n has a large enough number m of samples and a bounded corruption level, as specified by the following assumption, whose role we explain in the next paragraph.

Assumption 1. *(bounded corruption) Every D_n can be split in $2T$ batches, each having corruption level of at most $\epsilon < 0$.*

Robust OFTRL Guarantees. To prove the convergence of Algorithm 2, we first need to provide an upper bound on the quality of the solution that robust OFTRL (Algorithm 1) outputs. We show that after sufficiently many iterations T, robust OFTRL with gradient estimators defined in the previous section can find an approximate saddle point to (3), but with bounded domains $\mathcal{D}$ and $\mathcal{H}$. Similarly to (Prasad et al. 2020), to avoid statistical issues, we split the original dataset of m samples in $2T$ equal batches, assuming that each batch has corruption level at most $\epsilon < 0.5$. In each iteration, we apply each gradient estimator on a fresh batch. We refer to this process as *batch-splitting*.

Lemma 1. *There exists T such that the output of Algorithm 1 run for T iterations on $f = \mathcal{L}$, with $\mathcal{X} = \mathcal{D}$, $\mathcal{Y} = \mathcal{H}$, $g_{X,t} = \hat{g}_h$, $g_{Y,t} = \hat{g}_d$, and* batch-splitting*, satisfies*

$$\max_{d \in \mathcal{D}} \mathcal{L}(d, \bar{h}, M_n) - \min_{h \in \mathcal{H}} \mathcal{L}(\bar{d}, h, M_n) \leq 7C(\delta) \quad (4)$$

under Assumption 1, with probability at least $1-\delta$, where $C(\delta) := \sqrt{S} \left(E_1\left(\frac{m}{2T}, \epsilon, \frac{\delta}{T}\right) h_{max} + \frac{\sqrt{A} E_2\left(\frac{m}{2T}, \epsilon, \frac{\delta}{T}\right)}{1-\gamma} \right)$.

Now, we want to provide a bound on the quality of the output of robust OFTRL w.r.t. the true solution of (3). Consider the set of $d \geq 0$ that satisfy the Bellman flow constraint in M_n and denote its *Hoffman constant* (Garber 2019) by σ_n. To simplify the exposition, we define quantity $\alpha(M_n, \delta)$:

$$\alpha(M_n, \delta) = \sqrt{\frac{14C(\delta)}{\lambda}} + C_n' + \sqrt{2C_n' \left(\frac{\|r_n\|}{\lambda} + \frac{1}{1-\gamma} \right)}$$

where $C_n' = \frac{\|r_n\|_2 \sqrt{S} \sigma_n^{-1/2}}{(1-\gamma)h_{max}}$. For a generic MDP $M(\pi)$, we analogously define $\alpha(M(\pi), \delta)$. Next we show that robust OFTRL outputs an approximately optimal solution to (3).

Theorem 4. *Consider the robust OFTRL from Lemma 1, and assume its number of iterations T is s.t.* (4) *is holds. Under Assumption 1, the output of robust OFTRL $\bar{d}$ satisfies $\|d_n^* - \bar{d}\|_2 \leq \alpha(M_n, \delta)$ with probability at least $1-\delta$.*

Convergence of Algorithm 2. Now, we are ready to derive convergence guarantees of Algorithm 2. To do so, we need two additional assumptions, which we use to establish contraction properties of repeated retraining. The first one, ϵ-sensitivity is a standard in the literature on performative prediction, and we take it from prior work on performative RL (Mandal, Triantafyllou, and Radanovic 2023).

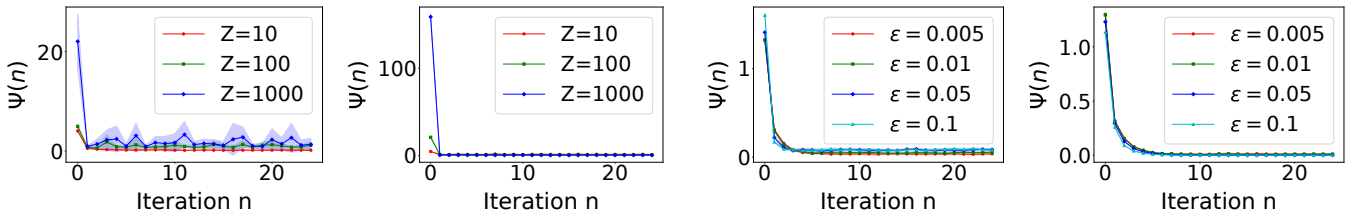
Assumption 2. *($\tilde{\epsilon}$-sensitivity) For any two MDPs $M(\pi)$ and $M(\pi')$, the following holds $\|r^\pi - r^{\pi'}\|_2 \leq \tilde{\epsilon}_r \|d^\pi - d^{\pi'}\|_2$ and $\|P^\pi - P^{\pi'}\|_2 \leq \tilde{\epsilon}_p \|d^\pi - d^{\pi'}\|_2$.*

The second one, is a rather weak assumption requiring that any MDP induced by the deployed policy does not have an infinite factor α for fixed δ.

Assumption 3. *($\bar{\alpha}$-boundedness) For any MDP $M(\pi)$ induced by stationary policy π we have $\alpha(M(\pi), \delta) \leq \bar{\alpha}(\delta)$.*

Note that after each round n, the distance between d_n and d_n^* is at most $\bar{C}(\delta) := \bar{\alpha}(\delta) + c\sqrt{SA}$. This is due to the definition of $\tilde{d}_n$—we obtain $\tilde{d}$ from $\bar{d}$ by adding a constant $c > 0$ to each of its entry — and the robust OFTRL guarantees for $\bar{d}$. Using this bound and the previous two assumptions we derive a convergence result for Algorithm 2.

Theorem 5. *(Informal Statement) Under Assumption 1, Assumption 2 and Assumption 3, there exist λ and N such that the output of Algorithm 2 satisfies $\tilde{d}_N \in \{d \in \mathcal{D} : \|d - d_S\|_2 \leq \tilde{C} := 4 \cdot \bar{C}(\delta/N)\}$ with probability at least $1-\delta$, where d_S is a performatively stable policy.*



(a) Fixed ϵ, naive (b) Fixed ϵ, robust (c) Fixed Z, naive (d) Fixed Z, robust

Figure 1: Convergence of repeated retraining approaches: (a) and (d) consider a non-robust variant of Algorithm 3 that utilizes naive gradient averaging; (b) and (c) consider Algorithm 3 that utilizes robust gradient estimation. The y axis is the normalised distance between consecutive repeated retraining solutions $\Psi(n) := c_n \cdot ||\tilde{d}_{n+1} - \tilde{d}_n||$ where $c_n = 1/\|\tilde{d}_n\|_2$, similar to the experiments in (Mandal, Triantafyllou, and Radanovic 2023), while the x axis is the number of repeated retraining iterations. For Fig. 1a and Fig. 1b, $\epsilon = 0.01$ and we vary Z. For Fig. 1c and Fig. 1d, $Z = 15$ and we vary ϵ. Each line presents the average over 10 experiments initialised with different seeds.

Theorem 5 states that Robust Repeated Retraining (Algorithm 2) exhibits last-iterate convergence to an approximate stable point. Ignoring ϵ-independent additive terms, the approximation error is proportional $\sqrt{\epsilon}$. Namely, factor $\bar{C}(\delta/N)$ depends on $\bar{\alpha}(\delta)$, which has square root dependence on ϵ since C is linear in ϵ. In prior work on corruption-robust offline RL, suboptimality gaps can scale as $O(\sqrt{\epsilon})$ or $O(\epsilon)$, depending on the exact setting and assumption (e.g., see (Zhang et al. 2022; Nika et al. 2024)). However, we note that our setting is not directly comparable, since it combines offline RL with repeated retraining. We leave the analysis of the tightness of this bound for future work.

The result in Theorem 5 provides an asymptotic convergence guarantee w.r.t. λ and N. Following the proof of Theorem 1 from (Mandal, Triantafyllou, and Radanovic 2023), it is easy to show that $\lambda > \lambda_0$ and $N > \frac{1}{1-\lambda_0} \cdot \log(\frac{2}{\bar{C}_0 \cdot (1-\gamma)})$ guarantee the convergence.[1] Here, $\lambda_0 = \frac{24 S^{3/2}(2\bar{\epsilon}_r + 5\bar{\epsilon}_p)}{(1-\gamma^4)}$, while $\bar{C}_0$ is a lower bound on $4 \cdot \bar{C}(\delta/N)$ independent of N, but possibly dependent on λ.

We can further assess the return that policy defined by $\tilde{d}_N$, i.e., $\pi^{\downarrow \tilde{d}_N}$, achieves in $\mathcal{M}(\pi_N)$. Assuming that the initial state distribution has a full support over state space, one can show that this return is comparable to the return of a performatively stable policy π_S in $M(\pi_S)$: it is worse by at most an instance-specific constant proportional to $\tilde{C}$.

Experimental Evaluation

In this section we experimentally test the efficacy of our approach in performative RL under corruption.

Environment. Our MDP model is a $W \times W$ gridworld environment, inspired by the gridworld environment in (Triantafyllou, Singla, and Radanovic 2021), where state $s = (i, j)$ encodes the location/cell that the agent occupies. In round n, the reward function is defined as $r_n(s, a) = R_{i,j} - c_p \cdot \sum_{a \in \mathcal{A}} d_n(i \cdot W + j, a)$. We set $W = 8$, while the values $R_{i,j}$ are defined as in the gridworld environment used in (Triantafyllou, Singla, and Radanovic 2021). The transitions are deterministic and only controlled by the four agent's actions (left, right, up, down). Samples are transition tuples (s_i, s'_i, a_i, r_i). In corrupted samples we add Gaussian noise $N(Z, 0.5)$ to r_i and we replace s'_i with a random state s' with probability exponentially decreasing with the distance between s' and s'_i on the grid. We study the convergence of the robust repeated retraining based on OFTRL (Algorithm 2). As a baseline approach, we consider a version of repeated retraining based on OFTRL which uses a naive estimator of g_d instead of Algorithm 3. In all the experiments, we set $\gamma = 0.99, c_p = 1$ and $\lambda = 0.001$. To create transition samples, we collect 1000 trajectories with an effective horizon of $1/(1-\gamma) = 100$.

Results. The plots in Fig. 1 show the convergence results for different values of the noise magnitude Z and the corruption frequency ϵ. We observe that naive gradient estimation results in more noisy convergence of repeated retraining, compared to robust gradient estimation. The effect is stronger when we fix ϵ and progressively increase Z. Then the curve of repeated retraining with a naive estimator oscillates with magnitude scaling with Z, while the curve of repeated retraining with a robust estimator stays virtually unaffected. Fixing Z and progressively increasing ϵ we see that both robust and naive retraining are affected, with the error increasing with ϵ. All these effects are in agreement with our theoretical results and they showcase the utility of robust gradient estimation.

Conclusion

We considered performative reinforcement learning under corrupted data. We introduced a repeated retraining approach and showed that it converges to an approximately stable policy, where the approximation error depends on the level of corruption. One of the most interesting future research directions is to investigate the tightness of the approximation error. Extending this work to RL settings with function approximation is another important avenue for future research.

[1] This lower bound on λ is the same as the one in Theorem 3 from (Mandal, Triantafyllou, and Radanovic 2023), which considers the finite sample case. The lower bounds on N are comparable but not the same, due to differences in the convergence criteria.

Acknowledgements

This work has been partially supported by project MIS 5154714 of the National Recovery and Resilience Plan Greece 2.0 funded by the European Union under the NextGenerationEU Program. This research was, in part, funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project number 467367360.

References

Ahn, K.; Jain, P.; Ji, Z.; Kale, S.; Netrapalli, P.; and Shamir, G. I. 2022. Reproducibility in optimization: Theoretical framework and limits. *Advances in Neural Information Processing Systems*, 35: 18022–18033.

Beznosikov, A.; Sadiev, A.; and Gasnikov, A. 2020. Gradient-free methods with inexact oracle for convex-concave stochastic saddle-point problem. In *International Conference on Mathematical Optimization Theory and Operations Research*, 105–119. Springer.

Brown, G.; Hod, S.; and Kalemaj, I. 2022. Performative prediction in a stateful world. In *International conference on artificial intelligence and statistics*, 6045–6061. PMLR.

d'Aspremont, A. 2008. Smooth optimization with approximate gradient. *SIAM Journal on Optimization*, 19(3): 1171–1183.

Devolder, O. 2013. *Exactness, inexactness and stochasticity in first-order methods for large-scale convex optimization*. Ph.D. thesis, CORE UCLouvain Louvain-la-Neuve, Belgium.

Devolder, O.; Glineur, F.; and Nesterov, Y. 2014. First-order methods of smooth convex optimization with inexact oracle. *Mathematical Programming*, 146: 37–75.

Diakonikolas, I.; Kane, D. M.; and Pensia, A. 2020. Outlier robust mean estimation with subgaussian rates via stability. *Advances in Neural Information Processing Systems*, 33: 1830–1840.

Dvinskikh, D.; Tominin, V.; Tominin, Y.; and Gasnikov, A. 2022. Gradient-free optimization for non-smooth minimax problems with maximum value of adversarial noise. *arXiv preprint arXiv*, 2202.

Garber, D. 2019. Logarithmic regret for online gradient descent beyond strong convexity. In *The 22nd International Conference on Artificial Intelligence and Statistics*, 295–303. PMLR.

Hardt, M.; and Mendler-Dünner, C. 2023. Performative prediction: Past and future. *arXiv preprint arXiv:2310.16608*.

Huang, K.; and Zhang, S. 2022. New first-order algorithms for stochastic variational inequalities. *SIAM Journal on Optimization*, 32(4): 2745–2772.

Izzo, Z.; Ying, L.; and Zou, J. 2021. How to learn when data reacts to your model: performative gradient descent. In *International Conference on Machine Learning*, 4641–4650. PMLR.

Juditsky, A.; Nemirovski, A.; and Tauvel, C. 2011. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1): 17–58.

Korpelevich, G. M. 1976. The extragradient method for finding saddle points and other problems. *Matecon*, 12: 747–756.

Letchford, J.; MacDermed, L.; Conitzer, V.; Parr, R.; and Isbell, C. 2012. Computing optimal strategies to commit to in stochastic games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 26, 1380–1386.

Li, Q.; and Wai, H.-T. 2022. State dependent performative prediction with stochastic approximation. In *International Conference on Artificial Intelligence and Statistics*, 3164–3186. PMLR.

Lu, S. 2023. Bilevel optimization with coupled decision-dependent distributions. In *International Conference on Machine Learning*, 22758–22789. PMLR.

Mandal, D.; Nika, A.; Kamalaruban, P.; Singla, A.; and Radanović, G. 2024. Corruption Robust Offline Reinforcement Learning with Human Feedback. *arXiv preprint arXiv:2402.06734*.

Mandal, D.; Triantafyllou, S.; and Radanovic, G. 2023. Performative reinforcement learning. In *International Conference on Machine Learning*, 23642–23680. PMLR.

Mendler-Dünner, C.; Perdomo, J.; Zrnic, T.; and Hardt, M. 2020. Stochastic optimization for performative prediction. *Advances in Neural Information Processing Systems*, 33: 4929–4939.

Miller, J. P.; Perdomo, J. C.; and Zrnic, T. 2021. Outside the echo chamber: Optimizing the performative risk. In *International Conference on Machine Learning*, 7710–7720. PMLR.

Mokhtari, A.; Ozdaglar, A.; and Pattathil, S. 2019. Proximal point approximations achieving a convergence rate of o (1/k) for smooth convex-concave saddle point problems: Optimistic gradient and extra-gradient methods. *arXiv preprint arXiv:1906.01115*, 3.

Mokhtari, A.; Ozdaglar, A.; and Pattathil, S. 2020. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. In *International Conference on Artificial Intelligence and Statistics*, 1497–1507. PMLR.

Narang, A.; Faulkner, E.; Drusvyatskiy, D.; Fazel, M.; and Ratliff, L. J. 2023. Multiplayer performative prediction: Learning in decision-dependent games. *Journal of Machine Learning Research*, 24(202): 1–56.

Nedić, A.; and Ozdaglar, A. 2009. Subgradient methods for saddle-point problems. *Journal of optimization theory and applications*, 142: 205–228.

Neff, G. 2016. Talking to bots: Symbiotic agency and the case of Tay. *International Journal of Communication*.

Nemirovski, A. 2004. Prox-method with rate of convergence O (1/t) for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1): 229–251.

Nesterov, Y. 2007. Dual extrapolation and its applications to solving variational inequalities and related problems. *Mathematical Programming*, 109(2-3): 319–344.

Nika, A.; Mandal, D.; Singla, A.; and Radanovic, G. 2024. Corruption-Robust Offline Two-Player Zero-Sum Markov Games. In *International Conference on Artificial Intelligence and Statistics*, 1243–1251. PMLR.

Orabona, F. 2019. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*.

Ouyang, Y.; and Xu, Y. 2021. Lower complexity bounds of first-order methods for convex-concave bilinear saddle-point problems. *Mathematical Programming*, 185(1-2): 1–35.

Perdomo, J.; Zrnic, T.; Mendler-Dünner, C.; and Hardt, M. 2020. Performative prediction. In *International Conference on Machine Learning*, 7599–7609. PMLR.

Piliouras, G.; and Yu, F.-Y. 2023. Multi-agent performative prediction: From global stability and optimality to chaos. In *Proceedings of the 24th ACM Conference on Economics and Computation*, 1047–1074.

Pollatos, V.; Mandal, D.; and Radanovic, G. 2024. On Corruption-Robustness in Performative Reinforcement Learning. *Extended Version; To Appear on arXiv*.

Polyak, B. T. 1987. Introduction to optimization. optimization software. *Inc., Publications Division, New York*, 1: 32.

Prasad, A.; Suggala, A. S.; Balakrishnan, S.; and Ravikumar, P. 2020. Robust estimation via robust gradient estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(3): 601–627.

Rank, B.; Triantafyllou, S.; Mandal, D.; and Radanovic, G. 2024. Performative Reinforcement Learning in Gradually Shifting Environments. In *The 40th Conference on Uncertainty in Artificial Intelligence*.

Robinson, J. 1951. An iterative method of solving a game. *Annals of mathematics*, 296–301.

Triantafyllou, S.; Singla, A.; and Radanovic, G. 2021. On blame attribution for accountable multi-agent sequential decision making. *Advances in Neural Information Processing Systems*, 34: 15774–15786.

Tseng, P. 1995. On linear convergence of iterative methods for the variational inequality problem. *Journal of Computational and Applied Mathematics*, 60(1-2): 237–252.

Tseng, P. 2008. On accelerated proximal gradient methods for convex-concave optimization. *submitted to SIAM Journal on Optimization*, 2(3).

Wang, Y.; Mianjy, P.; and Arora, R. 2021. Robust Learning for Data Poisoning Attacks. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 10859–10869. PMLR.

Wei, C.-Y.; Lee, C.-W.; Zhang, M.; and Luo, H. 2021. Linear Last-iterate Convergence in Constrained Saddle-point Optimization. In *International Conference on Learning Representations*.

Wu, F.; Li, L.; Xu, C.; Zhang, H.; Kailkhura, B.; Kenthapadi, K.; Zhao, D.; and Li, B. 2022. Copa: Certifying robust policies for offline reinforcement learning against poisoning attacks. *arXiv preprint arXiv:2203.08398*.

Yan, W.; and Cao, X. 2024. Zero-regret performative prediction under inequality constraints. *Advances in Neural Information Processing Systems*, 36.

Ye, C.; Yang, R.; Gu, Q.; and Zhang, T. 2024. Corruption-robust offline reinforcement learning with general function approximation. *Advances in Neural Information Processing Systems*, 36.

Zhan, W.; Huang, B.; Huang, A.; Jiang, N.; and Lee, J. 2022. Offline reinforcement learning with realizability and single-policy concentrability. In *Conference on Learning Theory*, 2730–2775. PMLR.

Zhang, L.; Yang, J.; Karbasi, A.; and He, N. 2024. Optimal guarantees for algorithmic reproducibility and gradient complexity in convex optimization. *Advances in Neural Information Processing Systems*, 36.

Zhang, X.; Chen, Y.; Zhu, X.; and Sun, W. 2022. Corruption-robust offline reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, 5757–5773. PMLR.

Zhong, H.; Yang, Z.; Wang, Z.; and Jordan, M. I. 2021. Can Reinforcement Learning Find Stackelberg-Nash Equilibria in General-Sum Markov Games with Myopic Followers? *arXiv preprint arXiv:2112.13521*.